

WHAT IS CLAIMED IS:

1                   1.       A computer-implemented method of generating concept units from  
2 user search queries, the method comprising:  
3                   receiving a plurality of queries, each query comprising a string of one or more  
4 words;  
5                   tokenizing each query string to produce one or more tokens for each query,  
6 wherein said tokens for said queries form an initial set of units;  
7                   combining units from the initial set of units that appear adjacent each other in  
8 a query to form a second set of units;  
9                   validating the second set of units;  
10                  repeating the steps of combining and validating one or more times using the  
11 second set of units in place of the initial set of units until a convergence condition is satisfied,  
12 wherein a final set of units is formed once the convergence condition has been satisfied; and  
13                  storing the final set of units to a memory.

1                   2.       The method of claim 1, wherein receiving includes receiving one or  
2 more query log files, each query log file including a plurality of queries.

1                   3.       The method of claim 2, further comprising consolidating the plurality  
2 of queries from the one or more query log files into a single consolidated query file.

1                   4.       The method of claim 3, wherein consolidating includes removing  
2 duplicates of queries and incrementing a count associated with each individual query each  
3 time a duplicate of said individual query is removed, wherein the consolidated file includes a  
4 list of individual queries and counts associated therewith.

1                   5.       The method of claim 3, wherein the received query log files include  
2 query log files for each day of a week, and wherein consolidating includes forming a single  
3 consolidated query file including queries for the week.

1                   6.       The method of claim 1, further including generating unit extensions  
2 using the final set of units

1                   7.       The method of claim 6, wherein generating unit extensions includes  
2 identifying units that are subsets of other units.

1                    8.        The method of claim 6, further including storing the unit extensions to  
2 the memory.

1                    9.        The method of claim 1, further including generating unit associations  
2 using the final set of units.

1                    10.      The method of claim 9, wherein generating unit associations includes  
2 identifying units that are associated with other units.

1                    11.      The method of claim 10, further including storing the unit associations  
2 to the memory.

1                    12.      The method of claim 10, wherein identifying associated units includes  
2 determining which units appear in queries with other units.

1                    13.      The method of claim 1, further comprising generating unit alternatives  
2 after the convergence condition has been satisfied.

1                    14.      The method of claim 13, wherein generating unit alternatives includes  
2 determining whether an edit distance between two units in the final set of units is smaller than  
3 a threshold value, and if so, comparing the relative frequencies of the two units.

1                    15.      The method of claim 1, further comprising:  
2 generating unit extensions using the final set of units;  
3 generating unit associations using the final set of units; and  
4 generating unit alternatives using the final set of units.

1                    16.      The method of claim 15, further including storing the unit extensions,  
2 the unit associations and the unit alternatives to the memory.

1                    17.      The method of claim 15, wherein generating unit extensions includes  
2 identifying units that are subsets of other units, wherein generating unit associations includes  
3 identifying units that are associated with other units, and wherein generating unit alternatives  
4 includes determining whether an edit distance between two units in the final set of units is  
5 smaller than a threshold value, and if so, comparing the relative frequencies of the two units.

1                   18.     The method of claim 1, wherein validating includes for each combined  
2 unit in the second set of units, comparing a frequency of occurrence of the combined unit  
3 with a frequency of occurrence of each constituent unit in the combined unit.

1                   19.     The method of claim 1, wherein the convergence condition includes a  
2 threshold value, wherein the convergence condition is satisfied if a change in the number of  
3 units in the two second set of units between successive steps of combining and validating is  
4 smaller than or equal to the threshold value.

1                   20.     The method of claim 1, further including:  
2 receiving an individual query from a user;  
3 identifying one or more units in the individual query; and  
4 determining one or more suggestions to provide to the user responsive to the  
5 query using one or more of the unit extensions, unit associations and unit alternatives stored  
6 in the memory in association with the one or more units identified in the individual query.

1                   21.     A system for generating concept units from user search queries, the  
2 system comprising:  
3 a memory unit; and  
4 a processing module configured to receive one or more query log files, each  
5 query log file including a plurality of queries, each query including a string of one or more  
6 words, and wherein the processing module is further configured to:  
7 tokenize each query from the query log files to produce an initial set of units;  
8 and thereafter, iteratively, until a convergence condition is satisfied:  
9 combine units from the initial set of units that appear adjacent each  
10 other in a query to form a second set of units; and  
11 validate the second set of units, wherein the second set of units is used  
12 for each iteration; and  
13 once the convergence condition has been satisfied, store a final set of units to  
14 the memory unit.

1                   22.     The system of claim 21, further including one or more query log file  
2 sources for providing the query log files.

1                   23.     The system of claim 21, wherein the processing module is further  
2 configured to:  
3                   generate unit extensions using the final set of units;  
4                   generate unit associations using the final set of units;  
5                   generate unit alternatives using the final set of units; and  
6                   store the unit extensions, unit associations and unit alternatives to the memory  
7 unit in association with the final set of units.

1                   24.     The system of claim 21, wherein the received query log files include  
2 query log files for each day of a week, and wherein the processing module is further  
3 configured to consolidate the query log files into a single consolidated query file consisting of  
4 queries for the week.

1                   25.     The system of claim 24, wherein the processing module consolidates  
2 by removing duplicates of queries and incrementing a count associated with each individual  
3 query each time a duplicate of said individual query is removed, wherein the consolidated file  
4 includes a list of individual queries and counts associated therewith.

1                   26.     The system of claim 25, wherein the processing module determines a  
2 frequency of occurrence for each unit using the counts associated with the queries, and  
3 wherein the processing modules stores the unit frequencies to the memory unit in association  
4 with the final set of units.

1                   27.     The system of claim 21, wherein the memory unit and processing  
2 module are implemented in a search server device in a network.

1                   28.     A computer readable medium including code for causing a processor to  
2 generate concept units from a plurality of user search queries, each query comprising a string  
3 of one or more words ,wherein the code includes instructions to:

4                   a) tokenize each query string to produce one or more tokens for each query,  
5 wherein said tokens for said queries form an initial set of units;

6                   b) combine units from the initial set of units that appear adjacent each other in  
7 a query to form a second set of units;

8                   c) validate the second set of units;

9                   d) repeat b) and c) one or more times using the second set of units in place of  
10 the initial set of units until a convergence condition is satisfied, wherein a final set of units is  
11 formed once the convergence condition has been satisfied; and  
12                   store the final set of units to a memory module.

1                   29.     The computer-readable medium of claim 28, wherein the code further  
2 includes instructions to:  
3                   generate unit extensions using the final set of units;  
4                   generate unit associations using the final set of units;  
5                   generate unit alternatives using the final set of units; and  
6                   store the unit extensions, unit associations and unit alternatives to the memory  
7 module in association with the final set of units.

1                   30.     The computer-readable medium of claim 29, wherein the instructions  
2 to generate unit extensions includes instructions to identify units that are subsets of other  
3 units, wherein the instructions to generate unit associations includes instructions to identify  
4 units that are associated with other units, and wherein the instructions to generate unit  
5 alternatives includes instructions to determine whether an edit distance between two units in  
6 the final set of units is smaller than a threshold value, and if so, compare the relative  
7 frequencies of the two units.

1                   31.     The method of claim 1, wherein each word comprises one or a  
2 plurality of alphanumeric characters.